Unsupervised learning

Clustering

→ Combinatorial Approaches
(Kmeans, K medoid)

(Hierarchical clustering
Agglomerative (SL, CL, GA)
Divisive

→ Bump hunting
(A priori Algorithm for Market
Basket Analysis)

→ Probabilistic approaches
(Mixture Models including GMMs
Gaussian
Mixture Models

+ EM Algorithm

→ Latent variable Models → Factor Analysis

→ Principal Component Analysis

→ Independent Component Analysis

One limitation with the GMM approach is that the prototypes are assumed to come from one of K MVNs with this connection being exclusive (a point cannot be generated as a combination)

→ A GMM can be understood as a latent variable model using K hidden variables representing a one hot encoding of the cluster identity

→ An extension would be to study the model resulting from a vector of real valued latent variables

We consider a vector of latent variables $z^{(i)} \in \mathbb{R}^k$ (not necessarily in $\{0,1\}$

and we will assume that those $z^{(i)}$ follow a Gaussian distribution

$$p(z^{(i)}) = \mathcal{N}(z^{(i)}; \mu_0, \Sigma_0)$$

Provided that our observations are continuous, we can further assume that those observations can be accurately modelled by a Gaussian distribution. Extending the idea of GMM, the Factor Analysis Model assumes that the $x^{(i)}$ have been generated by a family of MVNs with means that are determined by the hidden variables $z^{(i)}$

$z^{(i)} \in \mathbb{R}^K$

$$p(x^{(i)} \mid z^{(i)}, \theta) = \mathcal{N}(Wz^{(i)} + \mu, \Psi)$$

$x^{(i)} \in \mathbb{R}^D$

$W \in \mathbb{R}^{D \times K}$ (factor loading matrix)

Factor Analysis can be thought as a generalization of a GMM as since we constraint the $z^{(i)}$ to be binary vectors (one hot encodings), each $x^{(i)}$ can only be generated from one of the K Multivariate Normal Distributions respectively centered at w_1 to w_k and we recover the classical GMM.

Note $(*)$ implies $x^{(i)}$ can read as

$$x^{(i)} = W z^{(i)} + \mu + \varepsilon^{(i)} \qquad \varepsilon^{(i)} \sim \mathcal{N}(0, \Psi)$$

Main issue: factor loading matrix W is not uniquely identifiable. To see this, assume $\mu = 0$, $\mu_0 = 0$, $\Sigma_0 = I$, take any rotation matrix R (i.e such $RR^T = I$)

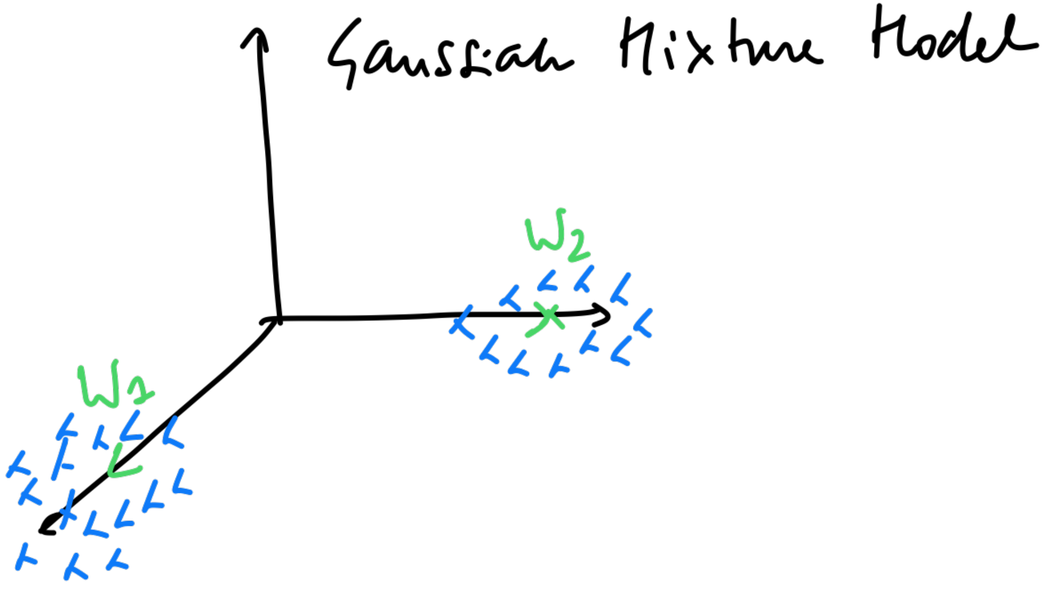
$$x = Wz + \varepsilon \qquad \mathbb{E}\{x\} = W\,\mathbb{E}\{z\} + \mathbb{E}\{\varepsilon\} = 0$$

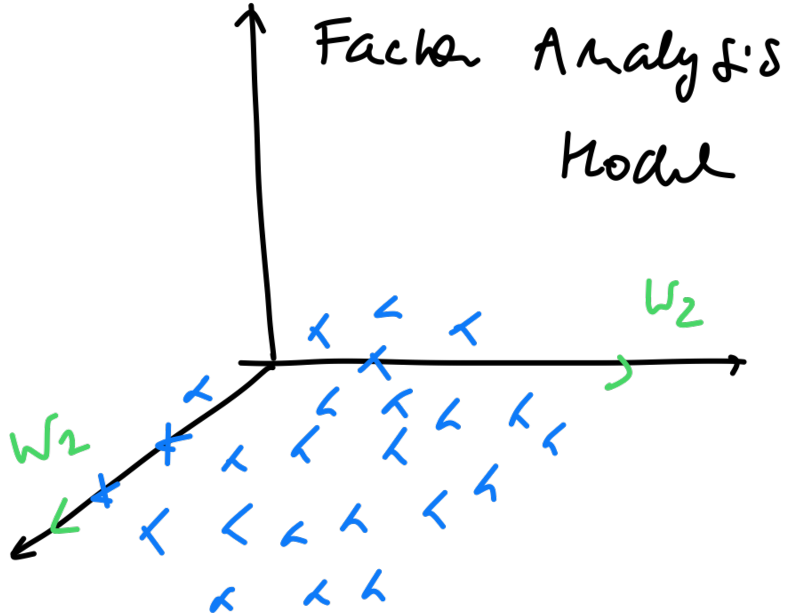
$$\tilde{x} = WRz + \varepsilon \qquad \mathbb{E}\{\tilde{x}\} = WR\,\mathbb{E}\{z\} + \mathbb{E}\{\varepsilon\} = 0$$

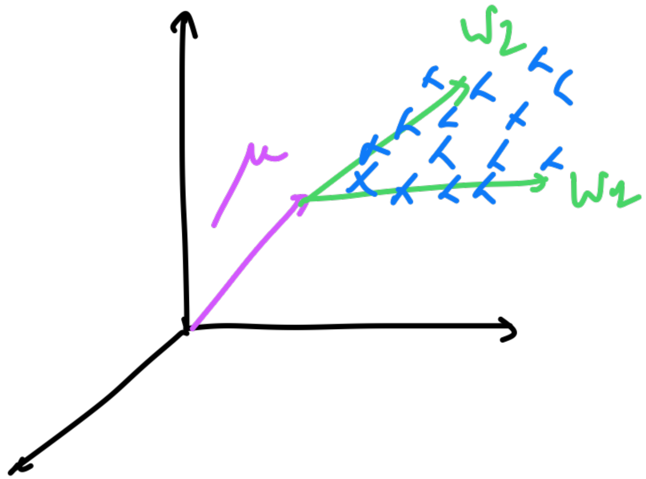
$$\mathrm{Cov}\{x\} = \mathbb{E}\{xx^T\} = W\,\mathbb{E}\{zz^T\}W^T + \mathbb{E}\{\varepsilon\varepsilon^T\} = WW^T + \Psi$$

$$\mathrm{Cov}\{\tilde{x}\} = \mathbb{E}\{\tilde{x}\tilde{x}^T\} = WR\,\mathbb{E}\{zz^T\}R^TW^T + \mathbb{E}\{\varepsilon\varepsilon^T\} = WRR^TW^T + \Psi$$

→ Conclusion: Two different factor loading matrices W, $\tilde{W} = WR$ lead to identical distributions.



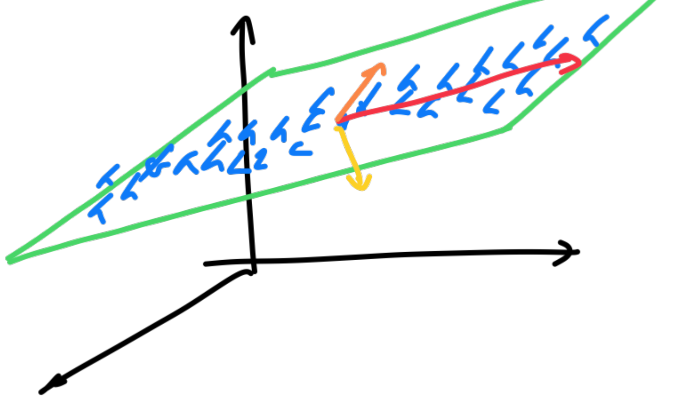




There exist several approaches in order to fix the ambiguity associated with the factor loading matrix :

1) Force W to have orthonormal columns

→ one of the cleanest solution to the identifiability problem is to force W to have orthonormal columns and sort these columns according to how much variance they capture in the data



$$W = [w_1, w_2, w_3]$$

$$w_1 \perp w_2 \perp w_3$$

2) Force W to be lower triangular (an approach used by the Bayesian community)

→ Main idea is to improve interpretability of the latent factors

→ first feature only generated by the first latent factor

second feature only generated by the first and second factors

...

on top of this the approach usually requires $W_{jj} > 0$

3) Sparsity promoting priors on W

→ can be achieved through regularization (e.g l_1)

instead of prespecifying which entries should be zero, we encourage the entries to vanish the approach is known as Sparse Factor Analysis

4) Choosing an informative rotation matrix

(Find R such that when applied to W, it improves interpretability)

5) Require the priors on the latent factors to be non Gaussian

→ can sometimes lead to unique identifiability of the factor loading matrix W

→ known as Independent Component Analysis (ICA)

#1 Orthonormal W : a.k.a Principal Component Analysis

→ Probabilistic intuition : take FA with $\mu_0 = 0$ $\Sigma_0 = I$

$\Psi = \sigma^2 I$ orthonormal W

→ solution through Max likelihood

Taking $\sigma^2 \to 0$ reduces the model to the classical PCA formulation (a.k.a Karhunen Loeve transform)

Theorem Consider a set of N prototype vectors $\{x^{(i)}\}_{i=1}^{N}$

$x^{(i)} \in \mathbb{R}^D$. We are looking for an orthonormal set of L basis vectors $w_j \in \mathbb{R}^D$ and their associated scores or latent factors $z^{(i)} \in \mathbb{R}^L$ such that we minimize the reconstruction error

$$J(W, Z) = \frac{1}{N} \sum_{i=1}^{N} \| x^{(i)} - \hat{x}^{(i)} \|^2 = \frac{1}{N} \sum_{i=1}^{N} \| x^{(i)} - W z^{(i)} \|^2$$

$W \in \mathbb{R}^{D \times L}$ $z^{(i)} \in \mathbb{R}^L$

(*) CLASSICAL PCA

Formulation (*) can equivalently read as

$$J(W, Z) = \| X - W Z^T \|_F^2$$

$$Z \in \mathbb{R}^{N \times L} \quad W \in \mathbb{R}^{D \times L} \quad X \in \mathbb{R}^{D \times N} \quad X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(N)} \\ | & | & & | \end{bmatrix}$$

$$\| A \|_F^2 = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} \qquad \text{for } A \in \mathbb{R}^{m \times n}$$

→ the optimal solution to the classical PCA problem is given by setting $\hat{W} = V_L$ where V_L encodes the eigenvectors corresponding to the largest eigenvalues of the empirical covariance matrix $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} (x^{(i)})^T$ after centering the $x^{(i)}$'s the optimal low dimensional representation of the data (given by the latent factors / scores $z^{(i)}$) can be obtained as $z^{(i)} = \hat{W}^T x^{(i)}$ (which is just the orthogonal projection of $x^{(i)}$ onto the latent space W)

As a result of the above theorem, to find the best dimension L representation of a set of N prototype vectors $\{x^{(i)}\}_{i=1}^{N}$, one can

1) Center the $x^{(i)}$ $\quad x^{(i)} \leftarrow x^{(i)} - \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$

2) Build the empirical covariance

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}(x^{(i)})^T$$

3) Compute the eigenvalue decomposition of

the covariance $\hat{\Sigma}$ and retain the largest L eigenvectors and set $\hat{W} = [v_1, \dots, v_L]$

4) Compute $z^{(i)}$ as $\hat{W}^T x^{(i)}$

5 : Independent Component Analysis

In PCA we required the factor loading matrix W to have orthonormal columns sorted according to how much variance they captured. In ICA we will still consider an orthonormal W but (as much of the ambiguity was coming from the spherical Gaussian which is rotation invariant) we will assume non Gaussian latent factors $z^{(i)}$

We still consider the general LVM $x = Wz$ (*)

In PCA we had $p(z^{(i)}) = \prod_{j=1}^{D} N(z_j^{(i)} \mid 0, I)$

In ICA, we still consider independence of $z_j^{(i)}$ but this time we forbid Gaussian priors

We end up with $p(z^{(i)}) = \prod_{j=1}^{D} p(z_j^{(i)})$

→ Popular approach: Maximization of the log likelihood through quasi second order method

→ implemented in FAST ICA. (see Scikit learn ICA)

→ Step 1 Derive an expression for $p(\{x^{(i)}\}_{i=1}^{N})$ as a function of the unknown Z, W

→ We will assume that our data has been whitened (see scikit learn 'whitening')

so that $\mathbb{E}\{x\} = 0$ $\mathbb{E}\{xx^T\} = I$

$$I = \mathbb{E}\{xx^T\} = \mathbb{E}\{WZZ^TW^T\} = WW^T$$

assuming $z_j^{(i)}$ are independent

$\Rightarrow$ W is orthonormal.

let us start with the cumulative distribution

letting $X = f(z) = Wz$ we can write

$$P_x(X \leq x) = P(f(z) \leq x) \underbrace{=}_{\substack{f \text{ linear} \\ \text{invertible}}} P_z(z \leq f^{-1}(x))$$

in order to write down the likelihood, we need the probability density function (pdf) of x which can be obtained by taking $\frac{d}{dx} P(X \leq x)$

$$\mathrm{pdf}_x = p(x) = \frac{d}{dx} P(X \le x) = \frac{d}{dx} P_z(Z \le f^{-1}(x))$$

$$= \frac{d}{dz} P_z(Z \le \underbrace{f^{-1}(x)}_{z}) \cdot \frac{dz}{dx}$$

$$= p(z) \left| \frac{dz}{dx} \right|$$

(to keep a non negative function we take the absolute value)

→ the multivariate equivalent is given

$$p(x) = p(z) \left| \det\left(\frac{\partial z}{\partial x} \right) \right|$$

in our case, since $X = Wz$ letting $V = W^{-1}$, we get

$$p(x) = p(z) . \left| \det(V) \right|$$

From the whitening of the data (which implied W orthonormal) we can write z as $z = Vx$

For any $x^{(i)}$ we have

$$p(x^{(i)}) = p(z^{(i)}) \, |\det(V)|$$

$$= p(Vx^{(i)}) \, |\det(V)|$$

From this, together with the independence assumption on the components of $z^{(i)}$, we can write our likelihood function as

$$\prod_{i=1}^{N} p(x^{(i)}) = \prod_{i=1}^{N} p(Vx^{(i)}) \, |\det(V)|$$

$$= \prod_{i=1}^{N} \prod_{j=2}^{L} p_{z_j}(v_j^T x^{(i)}) \, |\det(V)|$$

→ From the likelihood, taking the log and using the fact that V is orthonormal so that $\det(V) = \pm 1$ we get

$$\max_V \log \prod_{i=1}^{N} p(x^{(i)}) = \max_V \log \prod_{i=1}^{N} \prod_{j=2}^{L} p_{z_j}(v_j^T x^{(i)})$$

→ predefined non Gaussian priors are used for p_{z_j}

→ the resulting objective can be minimized through second order methods (such as in Fast ICA see scikit learn)